Early Education and Development, 25: 1118–1141

ISSN: 1040-9289 print/1556-6935 online
DOI: 10.1080/10409289.2014.904646

Routledge
Taylor & Francis Group

Validation of the Preschool Child Observation Record: Does It Pass the Test for Use in Head Start?

Katherine M. Barghaus and John W. Fantuzzo
Graduate School of Education, University of Pennsylvania

Research Findings: This study provides the first independent investigation of the second most widely used multidimensional assessment in Head Start—the Preschool Child Observation Record, Second Edition (COR-2). We conducted a comprehensive investigation into the validity of the COR-2 using data from all children in an urban school district's Head Start program ($N = 4{,}071$). Confirmatory factor analysis revealed a misfit between the 6 developer-defined categories and the data. Although exploratory analyses revealed a possible 4-factor solution, subsequent analyses indicated problems with this structure as well. Item response theory methods were used to determine whether there was support for the 5-point response scale of each item representing an appropriately sequenced set of skill points. Results indicated that nearly half of the COR-2 items had reversed or poorly spaced thresholds, suggesting potential problems with these items' functioning. *Practice or Policy:* Specific implications of the findings for the further development of the COR-2 in terms of its constructs and items as well as general implications for early childhood assessment are discussed.

Results from the Early Childhood Longitudinal Study's Birth cohort indicate that children from families living in poverty start kindergarten substantially behind more economically advantaged children in reading and mathematics (Denton Flanagan, McPhee, & Mulligan, 2009). The National Head Start program is the federal government's response to close these achievement gaps for children from low-income households by ensuring that these children are ready to start school. To meet its objectives, the Head Start program was developed based on, and continues to be guided by, developmental science theory and research. However, it was not until the Improving Head Start for School Readiness Act of 2007 was enacted that an explicit mandate was made to use scientific evidence to inform all aspects of the program (Zigler & Styfco, 2010). Clearly emphasized in this act was a call for the use of scientifically based assessment that must,

> (A) be developmentally, linguistically, and culturally appropriate for the population served; (B) be reviewed periodically, based on advances in the science of early childhood development; (C) be consistent with relevant, nationally recognized professional and technical standards related to the assessment of young children; (D) be valid and reliable in the language in which they are administered; (E) be administered by staff with appropriate training for such administration; (F) provide for appropriate accommodations for children with disabilities and children who are limited

Correspondence regarding this article should be addressed to Katherine M. Barghaus, Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104. E-mail: barghaus@upenn.edu

> English proficient; (G) be high-quality research-based measures that have been demonstrated to assist with the purposes for which they were devised. (Improving Head Start for School Readiness Act of 2007, Sec. 641A)

To help meet this call, the U.S. Congress charged the National Research Council (NRC) with developing guidelines to evaluate the quality of available early childhood assessments (NRC, 2008). Informed by developmental science, the NRC committee defined quality in terms of the capacity to measure important domains of children's functioning across time (i.e., cognitive, language, physical, social-emotional, and approaches toward learning) and sensitivity to unique child characteristics (e.g., sex, race, and language). The committee provided quality criteria for scientifically based assessment by drawing upon the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; herein referred to as the *Standards*). These quality criteria are organized into three categories—validity, reliability, and unbiasedness. Although each of these quality categories is important, the *Standards* notes that ''validity is . . . the most fundamental consideration in developing and evaluating tests'' (AERA, APA, & NCME, 1999, p. 9). Validity is essential because it refers to the extent to which there is evidence that test scores can be interpreted as intended (AERA, APA, & NCME, 1999).

According to the *Standards*, psychometrically sound assessments have validity evidence based on their content, response process, internal structure, relationships to other variables, and consequences of their use (AERA, APA, & NCME, 1999). Content validity evidence is derived from the systematic process used to develop and evaluate the targeted construct's definition and corresponding items (Downing & Haladyna, 1997; Kane, 2006a). Validity evidence based on the response process comes from documentation of the extent to which the assessment process is free from error (Downing, 2003). Response process validity for observational assessments, which are widely used with young children, is established through evidence that the observational process does not introduce error (Downing, 2003). Validity evidence based on the internal structure of an assessment refers to the extent to which the test and items conform to the targeted constructs and the intended use (AERA, APA, & NCME, 1999). Assessments that aim to capture multiple constructs should have evidence from factor analysis supporting their dimensionality (Gorsuch, 2003). In addition, if a measure aims to capture development, it is important to determine whether the items accurately reflect an ability continuum. Evidence based on an assessment's relationships to other variables is assessed by correlations between scores on the assessment and on other relevant measures (AERA, APA, & NCME, 1999). Finally, according to the *Standards*, validity evidence can also be provided based upon the consequences of using an assessment. This source of evidence is contentious because it indicates a broader definition of validity that includes both the interpretation of test scores and the consequences of their use (Cizek, 2012; Cook & Beckman, 2006; Kane, 2006b). Thus, validity experts have argued that the consequences of use be considered in concert with, but separate from, other aspects of validity evidence (e.g., Cizek, 2012).

To provide decision makers with information on the psychometric quality of assessments, the Administration for Children and Families commissioned a compendium of widely used measures (Halle, Zaslow, Wessel, Moodie, & Darling-Churchill, 2011). The compendium included formative assessments that (a) covered three or more of the domains of the Head

Start Child Outcomes Framework, (b) had some evidence base, and (c) were accessible for general use. Based on these criteria, only eight assessments were selected for review.[1] The purpose of this research was to use multivariate statistics and item response theory (IRT) to provide validity evidence for one of the eight measures. The Preschool Child Observation Record, Second Edition (COR-2), was selected because it is the second most widely used assessment in Head Start (Alkens et al., 2010), it corresponds to many states' early learning standards (Epstein, 2006; Epstein & Schweinhart, 2009), and it was used in one of the nation's largest and poorest school districts, which provides a nationally important context within which to examine this measure.

The COR was developed by the HighScope Educational Research Foundation to measure the learning and development of children 2.5 to 6 years old (HighScope, 1992). According to the developers, the COR may be used to monitor the progress of individual children and groups of children, to inform curriculum planning and instruction, and to assess the effectiveness of classrooms or programs (HighScope, 2010). In 1993, the first commercially available edition of the COR (COR-1) was released. It contained 30 items organized into six categories of development: (a) Initiative, (b) Social Relations, (c) Creative Representation, (d) Movement and Music, (e) Language and Literacy, and (f) Logic and Mathematics (HighScope, 1992). These categories corresponded to the five major domains of development (i.e., cognitive, language, physical, social-emotional, and approaches toward learning) that are nationally recognized as important for school readiness (NRC, 2008; Office of Head Start, 2010). Each item presented a continuum of five skill points for observers to rate children's level of skill development (HighScope, 1992). Observers used the categories, items, and skill points to classify and rate their observations of children's functioning (HighScope, 1992).

In 2003, HighScope released the second edition of the COR (COR-2), which differed from the COR-1 in several important ways. First, for every item the lowest skill point was changed from referring to a child ''not yet demonstrating'' a skill to exhibiting a ''basic exploration'' into a skill (HighScope, 2003). Second, COR-1 items were edited to reflect current literature on these key areas of development. New items were added to the Language and Literacy category and Logic and Mathematics category, which was renamed ''Mathematics and Science'' to reflect the changes (Neill, 2004). The final version of the COR-2 contained 32 items organized into the revised six categories (HighScope, 2010).

A comprehensive search for research on the COR-1 and COR-2 yielded one HighScope report and two published investigations of the psychometric properties of the COR-1 and one HighScope report on the COR-2. No detailed technical documentation on the content validity of the COR-2 was located. In contrast, some validity evidence for the response process was found for the COR-1 and COR-2. HighScope offers a recommended training program to provide instruction, practice, and feedback to support the accurate use of the measure. However, only limited empirical evidence was found on the consistency of the assessment process. For the COR-1, Epstein (1993) found high interobserver agreement with observers who received

[1]The eight assessments selected were the (a) Creative Curriculum Developmental Assessment; (b) Galileo Preschool Assessment Scales; (c) HighScope Child Observation Record; (d) Learning Accomplishment Profile–Third Edition; (e) Learning Accomplishment Profile–Diagnostic; (f) Learning Accomplishment Profile–Diagnostic, Spanish Edition; (g) Mullen Scales of Early Learning; and (h) Work Sampling System (Halle et al., 2011).

intensive training, whereas Schweinhart, McNair, Barnes, and Larner (1993) found moderate agreement with a sample of Head Start teachers who received less training. For the COR-2, moderate interobserver agreement, ranging from .69 to .79, has been found (HighScope, 2010).

The *Standards* notes that the type of analyses used to bring evidence to bear on the internal structure of an assessment depends on the intended use and interpretation of the assessment's scores (AERA, APA, & NCME, 1999). Key to the use and interpretation of the COR-1 and COR-2 is their ability to capture key school readiness domains and measure progress in each. Two studies examined the dimensionality of the COR-1 and one HighScope report examined this for the COR-2. Schweinhart et al. (1993) examined the psychometric properties of the COR-1 using 50 pairs of trained teaching teams to collect data on 484 children. A confirmatory factor analysis (CFA) indicated that the six categories did not fit the data well (goodness-of-fit index [GFI] = .79; GFI > .90 indicates a reasonable fit; Kline, 2005) and that the factors were highly correlated (range = .71–.86), suggesting that some may be redundant (Brown, 2006). Fantuzzo, Hightower, Grim, and Montes (2002) studied the validity of the COR-1 using data from 733 children enrolled in Head Start and 1,356 children from other preschool programs. Using exploratory factor analysis (EFA) and confirmatory cluster analysis, Fantuzzo et al. (2002) determined that a three-factor structure fit both samples (Cognitive Skills, Social Engagement, and Coordinated Movement). To date, the psychometric quality of the COR-2 has only been examined by HighScope (2010). In this study, Head Start teachers administered the COR-2 to 160 children in the spring and to 233 different children in the fall (HighScope, 2010). Based on principal component analysis, a complex four component structure with several items loading on multiple components was advanced.

Unfortunately, all three investigations of the dimensionality of the COR-1 and COR-2 ignored the categorical nature of the data and instead improperly treated them as continuous. Using standard factor analysis with categorical data is inappropriate because it violates foundational assumptions of the model and misrepresents the data (Flora & Curran, 2004). Furthermore, treating categorical data as continuous in factor analysis can yield a structure that distorts the underlying constructs and does not replicate across samples (Bernstein & Teng, 1989; McDermott et al., 2011). McDermott et al. (2011) noted that such results are nontrivial, as they promote interpretation and decision making based on spurious constructs. Advances have been made in psychometric science to address these issues, such as using polychoric correlations for factor analysis of categorical data (McDermott et al., 2011). A further issue with the examination of the dimensionality of the COR-2 is that it relied on principal component analysis. Snook and Gorsuch (1989) found that principal component analysis yields inflated component loadings and standard errors when there are less than approximately 40 items, which is the case for the COR-2. Finally, the examination of the COR-2 used a sample of approximately 200 children, which is half the size recommended by Gorsuch (2003) to ensure a viable structure.

The internal structure with respect to item functioning has also been examined for the COR-1. Fantuzzo et al. (2002) evaluated the five skill points of each COR-1 item to determine whether they represented a valid hierarchical developmental sequence. Using descriptive statistics, the researchers found that more than one third of the items had irregular distributions, suggesting that these items' skill points do not represent a progressive ability sequence. Currently, no published research has examined this important aspect of the skill points of the COR-2 items. Finally, studies of the COR-1 and COR-2 have investigated their validity based on relationships to other variables (Fantuzzo et al., 2002; Schweinhart et al., 1993; Sekino & Fantuzzo, 2005).

However, this research is not reviewed here given the shortcomings of the studies on the dimensionality and item skill points of the COR-1 and COR-2.

Despite its widespread use, there is no published, peer-reviewed research on the psychometric quality of the COR-2. The purpose of the present study was to start to bridge this knowledge gap by investigating the validity of the COR-2 for children attending preschool in the context of urban poverty. The dimensionality of the COR-2 was rigorously evaluated using a series of factor analyses. This process included both a CFA of the six-factor structure posited by the developers as well as a multistep exploratory and confirmatory examination of the internal structure. In addition, IRT methods were used to examine the five skill points of each item. This information was used to determine the extent to which the skill points correspond to a hierarchical developmental sequence.

METHOD

Participants

This study analyzed data from a larger study of a comprehensive early childhood educational program. The program consists of an evidence-based curriculum, a strong partnership with families, an evidence-based formative assessment, and professional development for teachers (Fantuzzo, Gadsden, & McDermott, 2010). Data for the present study included all children who were in a large, urban school district's Head Start program in 2006–2007. The analysis sample consisted of 4,071 children with COR-2 data for the fall, winter, and spring. Approximately half of the children in the sample were male, 5% were Caucasian, 70% were African American, 18% were Latino, 3% were Asian, 4% were other, and 5% were English language learners. The average age of the children was approximately 3.5 years old.

The school district in this study requires Head Start teachers to have a bachelor's degree and certification in early childhood education. The district also required the use of the COR-2 at the time to monitor children's learning and development in areas important for school readiness (the COR-1 was used before the COR-2). The COR-2 was completed three times each year so that it could inform lesson planning and provide information on the extent to which objectives and standards were being met. A subset of teachers was recruited to attend the 2-day training on the COR-2 recommended by HighScope, and they trained the other teachers at their school who had not attended the training (Waterman, McDermott, Fantuzzo, & Gadsden, 2012). Informal follow-up training was provided in subsequent professional development sessions (Waterman et al., 2012).

Measures

COR-2. The COR-2 is a widely used early childhood assessment of multiple domains of functioning important for school readiness (HighScope, 2010). It consists of 32 categorical items organized into six categories: Initiative, Social Relations, Creative Representation, Movement and Music, Language and Literacy, and Mathematics and Science (HighScope, 2010). Each of the 32 items contains five skill points ranging from less (1) to more (5) developed ability. Teachers record the highest skill point applicable for each item at a particular time. Skill points

are then averaged across all items within each category to create category scores and across all categories to create a total score.

Procedure

Testing the fit of the six COR-2 categories. A CFA of the purported COR-2 structure was performed using data from the fall for the 4,071 children with COR-2 data from the fall, winter, and spring. The goal of this analysis was solely to test the fit of the six COR-2 categories to the data. Gorsuch (2003) recommended using a sample of at least 400 to ensure stable correlations and a viable structure. The present study's sample was well in excess of this sample size guideline. Prior to performing the CFA, we performed an item analysis to look for items with severely constricted variance, to look for items with floor and ceiling effects, and to screen for potential data entry errors.

The raw data were used to estimate a categorical CFA model using Mplus 6.1 (Muthén & Muthén, 2010). Mean and variance adjusted robust weighted least squares was used for estimation. Flora and Curran (2004) showed that this method yields robust estimates with varying sample sizes, degrees of nonnormality, and levels of model complexity. Per Brown (2006), model fit was evaluated based on (a) several global goodness-of-fit indices, (b) the modification indices and completely standardized expected parameter change (SEPC) estimates to identify specific areas of model misfit, and (c) the practical and statistical interpretability of the model and its parameters. This three-pronged approach avoided the common mistake of solely using goodness-of-fit indices to evaluate fit (Brown, 2006). Specifically, ''the other two aspects of fit evaluation (specific areas of model misfit, parameter estimates) provide more specific information about the acceptability and utility of the solution'' (Brown, 2006, p. 113).

The global goodness-of-fit indices that were used as part of the model evaluation were the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the weighted root-mean-square residual (WRMR). Yu (2002) found that the WRMR performs well with nonnormal data and model misspecifications. The chi-square test was reported as well, but it was not relied on for decision making because of its sensitivity to large sample sizes (Brown, 2006). Simulation studies have found that good fit is characterized by CFI $\geq$ .950–.960, RMSEA $\leq$ .050–.060, and WRMR $\leq$ .950–1.000 (Hu & Bentler, 1999; Yu, 2002).

Global goodness-of-fit indices may indicate an adequate fit even when some observed relationships are not accounted for sufficiently (Brown, 2006). To identify areas of misfit, we inspected modification indices and SEPC estimates. Modification indices estimate the expected change in the chi-square statistic if a parameter not freely estimated in the model is freed (Brown, 2006). Given the sensitivity of chi-square to sample size, modification indices were considered in conjunction with SEPC estimates. SPEC estimates indicate the expected change in a parameter if it is freely estimated and therefore suggests whether respecification will lead to a statistical and a practically meaningful improvement (Brown, 2006). The results were inspected for SEPC values that suggested an item would load saliently ($\geq$.40) on a factor other than the one it was designated to load on a priori. Finally, the practical and statistical interpretability of the model—including the model specification; the size, significance, and direction of the factor loadings; and interfactor correlations—was used to evaluate fit (Brown, 2006). Robust model specification is indicated by having factors with four or more saliently loading items (Gorsuch, 2003).

To summarize, the fit of the six-factor CFA model was evaluated using the following criteria: (a) CFI $\geq$ .950–.960, RMSEA $\leq$ .050–.060, WRMR $\leq$ .950–1.000 (Hu & Bentler, 1999; Yu, 2002); (b) modification indices and SEPC estimates indicating areas of model misfit (Brown, 2006); and (c) the practical and statistical interpretability of the model (e.g., robust model specification, significant loadings, and reasonable interfactor correlations; Brown, 2006).

EFAs. The sample was randomly divided into two subsamples of approximately 2,035 children: (a) an exploratory sample for the EFAs and (b) a reserve sample for the confirmatory analyses of the factor structure derived from the EFAs. This two-step approach is recommended by factor analysis experts (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999) because the optimal factor structure is empirically uncovered using EFA, and then its fit to the data is cross-validated with CFA. Such an approach is especially appropriate when there is no or limited prior empirical and theoretical evidence to support using CFA (Brown & Moore, 2012).

Two-stage maximum likelihood estimation was used to calculate a polychoric correlation matrix in MicroFACT 2.0 for the exploratory sample (Waller, 2001). Per Knol and Berger (1988), the matrix was smoothed to reduce the number of Heywood cases (i.e., communalities $\geq$1) and to ensure positive semidefiniteness. To determine the initial number of factors to extract, we performed minimum average partialing (MAP) on the smoothed matrix (Velicer, 1976). The matrix was then used for iterative common factoring in SAS 9.3 using squared multiple correlations as initial communality estimates (McDermott et al., 2011). The analyses used varimax, equamax, and promax rotational procedures. The optimal structure was the one that met the criteria used by McDermott et al. (2011) for this type of analysis: (a) maximizes hyperplane count and item coverage indicating an approximation of simple structure (Yates, 1987), (b) produces the smallest root-mean-square residual (RMSR) and largest GFI (Waller, 2001), (c) has at least four salient items (loadings $\geq$ .40) per factor, (d) results in internally consistent factors ($r \geq .70$), and (e) yields an uncomplicated structure aligned with theory and research (Fabrigar et al., 1999).

CFAs. To confirm the optimal structure derived from the exploratory analyses, we submitted the reserve sample data to categorical CFA using the procedures specified previously. In addition, a higher order model was fit to provide a test for a general developmental status factor commonly found with many performance assessments (Gorsuch, 2003). The Schmid–Leiman transformation (Schmid & Leiman, 1957) is helpful in understanding higher order models by producing orthogonal first- and second-order factors and estimates of their unique contributions to explaining variance (Watkins, Wilson, Kotz, Carbone, & Babula, 2006). Per Brown (2006), an orthogonalized (i.e., applying the Schmid–Leiman transformation) higher order model was estimated by specifying the number of second- and first-order factors suggested by theoretical and empirical evidence. Item loadings on the second-order factor were estimated by extracting the variance this factor explained (Brown, 2006). The first-order factor loadings were residualized of the variance explained by the second-order factor, leaving only the variance uniquely accounted for by the first-order factors (Watkins et al., 2006).

Developmental sequence. The data were analyzed to understand the functioning of the skill points of each item in terms of the ordering and spacing of the estimated parameters. To do this, we estimated thresholds corresponding to the levels of the latent trait that separated two adjacent skill points (Andrich, 2010). These parameters should increase such that the

threshold for the boundary between Skill Points 1 and 2 is less than the threshold for Skill Points 2 and 3 (Bond & Fox, 2007). Thresholds that did not progress in this manner were identified as disordered or reversed (Bond & Fox, 2007). Thresholds were also inspected for reasonable spacing, which suggested that each skill point identified a unique level of the latent trait (Bond & Fox, 2007).

Threshold reversals have several potential causes, including incorrect skill point ordering, multidimensional responses, and skill points with low frequencies (Adams, Wu, & Wilson, 2012; Andrich, de Jong, & Sheridan, 1997). Given the multiple potential causes of reversals, their interpretation is debated. Some experts argue that threshold reversals indicate ''clear and unambiguous evidence of problems in the empirical ordering'' of the skill points (Andrich et al., 1997, p. 70). Others argue that reversals do not necessarily point to a problem with the skill point ordering but agree that they indicate that an item is malfunctioning (Adams et al., 2012). Both viewpoints do not recommend using remedial measures such as collapsing categories and instead suggest that items with threshold reversals need to be reviewed and revised (Adams et al., 2012; Bond & Fox, 2007).

In the present study, IRT methods were used to estimate the thresholds for each item. Both the partial credit model (PCM) and the generalized PCM (GPCM) were estimated using PARSCALE 4.1. To determine which model produced a better fit to the data, we compared the chi-squares and estimated internal consistency and information for each scale (du Toit, 2003; Embretson & Reise, 2000). Results from the selected model were used to detect reversed and potentially poorly spaced thresholds.

RESULTS

Item Analysis

On average, item responses were not skewed (M skewness = .05, range = −.40 to .73), but they were platykurtic (M kurtosis = −.94, range = −1.42 to −0.36). Each of the skill points of every item was used, although some infrequently. Floor and ceiling effects were investigated by examining the occurrence of extreme item response patterns (e.g., 1 or 5 on every item). About 1% of the children had extreme response patterns, with most receiving a 1 on every item. Covariate information was used to determine the likelihood that these patterns were due to data recording errors. In general, this information suggested that these responses were genuine (e.g., those who received all fives were typically older and did well on a validated measure of early academic achievement). Furthermore, there were only a few cases of extreme response patterns, and thus they were not removed from the data.

Testing the Fit of the Six COR-2 Categories

The CFA model specification of the six COR-2 categories is displayed in Figure 1. This specification indicated no double-loading items, uncorrelated measurement errors, correlated factors, and an overidentified model with $df = 449$. Data from the fall for the 4,071 children with COR-2 data at all three time points were used to estimate the model (see Table 1).

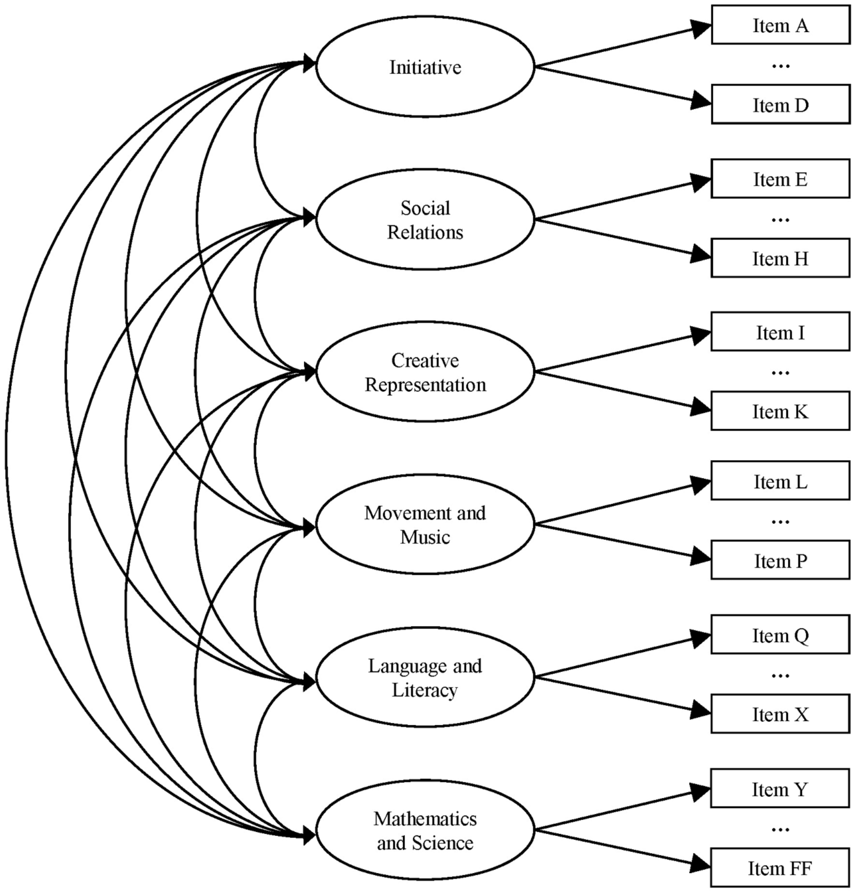

FIGURE 1 Confirmatory factor analysis model of six categories of the Preschool Child Observation Record, Second Edition. Note that this figure provides a simplified presentation and does not show all of the items (as indicated by the ellipses) or model parameters. Please refer to Table 1 to see the number and nature of items on each factor. Categories are from HighScope (2010).

For the six-factor model, the CFI was .978, suggesting a good fit, but at .066 and 2.602 the RMSEA and WRMR were higher than the criteria set for a good-fitting model ($\chi^2(449) = 8{,}498.46$, $p < .00001$). In addition, Brown (2006) noted that the CFI is more likely to suggest an acceptable fit than other indices because it compares the fit of the theoretical model to the fit of a model in which the items are unrelated. The modification indices, SEPC estimates, interfactor correlations, and factor specifications pointed to problems with this structure. The modification indices and SEPC estimates suggested that many items on the Social Relations, Initiative, and Creative Representation factors would load saliently ($\geq .40$) on more than one of these factors. This indicated that these three factors might have been better represented by one factor, a hypothesis supported by the high correlations between these factors ($M\ r = .96$).

As shown in Table 2, all six factors were highly correlated ($M\ r = .91$, range = .85–.97). A correlation of .91 indicates that only 17% of a factor's variance is unique, whereas 83% is

TABLE 1
Confirmatory Factor Analysis Results for the Six Categories of the Preschool Child Observation Record, Second Edition ($N = 4{,}071$)

Item	*Estimate*	*SE*	*Standardized estimate*	*SE*
1. Initiative				
A. Making choices and plans	1	0	0.85	0.01
B. Solving problems with materials	0.97	0.01	0.82	0.01
C. Initiating play	1.02	0.01	0.86	0.01
D. Taking care of personal needs	0.94	0.01	0.80	0.01
2. Social Relations				
E. Relating to adults	1	0	0.86	0.01
F. Relating to other children	0.97	0.01	0.84	0.01
G. Resolving interpersonal conflict	0.94	0.01	0.80	0.01
H. Understanding and expressing feelings	1.00	0.01	0.86	0.01
3. Creative Representation				
I. Making and building models	1	0	0.86	0.01
J. Drawing and painting pictures	0.98	0.01	0.84	0.01
K. Pretending	1.00	0.01	0.86	0.01
4. Movement and Music				
L. Moving in various ways	1	0	0.84	0.01
M. Moving with objects	0.94	0.01	0.79	0.01
N. Feeling and expressing steady beat	1.03	0.01	0.86	0.01
O. Moving to music	1.01	0.01	0.85	0.01
P. Singing	1.00	0.01	0.85	0.01
5. Language and Literacy				
Q. Listening to and understanding speech	1	0	0.89	0.00
R. Using vocabulary	0.97	0.01	0.87	0.01
S. Using complex patterns of speech	0.98	0.01	0.87	0.01
T. Showing awareness of sounds in words	0.96	0.01	0.86	0.01
U. Demonstrating knowledge about books	0.95	0.01	0.85	0.01
V. Using letter names and sounds	0.90	0.01	0.80	0.01
W. Reading	0.89	0.01	0.80	0.01
X. Writing	0.92	0.01	0.82	0.01
6. Mathematics and Science				
Y. Sorting objects	1	0	0.87	0.01
Z. Identifying patterns	0.97	0.01	0.84	0.01
AA. Comparing properties	1.02	0.01	0.89	0.00
BB. Counting	0.95	0.01	0.83	0.01
CC. Identifying position and direction	0.98	0.01	0.86	0.01
DD. Identifying sequence, change, and causality	1.03	0.01	0.90	0.00
EE. Identifying materials and properties	0.98	0.01	0.86	0.01
FF. Identifying natural and living things	0.95	0.01	0.83	0.01

Note. In the unstandardized model, one indicator per factor was fixed to 1 to define the factor metric (Brown, 2006). Categories and item titles are from HighScope (2010).

redundant. Many researchers have noted that high interfactor correlations (e.g., greater than .80–.85) provide ‘‘strong evidence to question’’ the existence of distinct constructs (Brown & Moore, 2012, p. 373). Williams, Ford, and Nguyen (2002) noted that ‘‘most researchers would not want to attempt the argument that two factors with such a high correlation [referring to a

TABLE 2
Interfactor Correlations for the Six Categories of the Preschool Child Observation Record, Second Edition

Category	*1*	*2*	*3*	*4*	*5*	*6*
1. Initiative	—	.97	.96	.89	.93	.90
2. Social Relations		—	.94	.86	.92	.88
3. Creative Representation			—	.89	.93	.91
4. Movement and Music				—	.85	.86
5. Language and Literacy					—	.95
6. Mathematics and Science						—

Note. Categories are from HighScope (2010).

correlation of .86] are meaningfully different'' (p. 373). Thus, the high interfactor correlations suggested that a model with fewer factors might fit better.

Finally, the six-factor model had one factor with just three items. This specification is problematic because experts have noted that to robustly define a dimension four or more indicators are needed (Gorsuch, 2003). In sum, the WRMR, changes suggested by the modification and SEPC estimates, high interfactor correlations, and model specifications problems all supported the conclusion that the six-factor model did not fit the data well. Using CFA may have been premature, as it relies on a robust theoretical and empirical foundation to determine the appropriate number of factors (Brown & Moore, 2012, p. 373). Thus, EFA and CFA were used to empirically determine the optimal factor structure of the COR-2.

EFAs

The fall data on the 4,071 children were randomly divided into an exploratory ($n = 2{,}036$) and a confirmatory ($n = 2{,}035$) sample. Two-stage maximum likelihood estimation was used to calculate a smoothed polychoric correlation matrix for the exploratory sample (Knol & Berger, 1988; Waller, 2001). The smoothed correlation matrix was submitted to MAP, which suggested four potentially viable factors, and therefore two- through six-factor solutions were evaluated.

The four-factor promax ($k = 4$) model with initial equamax rotation was determined to be the optimal solution, as it met all of the evaluation criteria. The six- and five-factor models produced factors with too few salient items, and the six-factor model did not reproduce the developer-defined categories. Solutions with less than four factors generally reproduced the factors in the four-factor model in a more complex manner (e.g., more items loading on two factors, less coverage of the items, lower hyperplane count). The four-factor model jointly maximized hyperplane count and item coverage, providing the best approximation of simple structure. The fit indices suggested that this model provided a good fit to the data (GFI = .9995 and RMSR = .02). The four-factor model retained five or more salient items per factor, and all of the factors were internally consistent (range = .89–.96). A total of 31 of the 32 items were salient, with only Item U, ''Demonstrating knowledge about books,'' not loading saliently on any factor. Item R, ''Using vocabulary,'' loaded on two factors and was removed from both per Comrey's (1988) recommendation, resulting in 30 items being retained.

Table 3 provides the final factors and their component items and pattern loadings. These results indicated that the four-factor model met the final EFA evaluation criteria of making

TABLE 3
Equamax, Promax ($k = 4$) Rotated Factor Pattern Loadings for the Preschool Child Observation Record, Second Edition

Item	*1*	*2*	*3*	*4*
1. Social Engagement				
E. Relating to adults	**.81**	−.09	.00	.14
F. Relating to other children	**.81**	.02	.03	.01
A. Making choices and plans	**.79**	.07	−.10	.11
G. Resolving interpersonal conflict	**.64**	.09	.01	.11
C. Initiating play	**.61**	.13	.26	−.09
S. Using complex patterns of speech	**.61**	−.05	−.03	.39
B. Solving problems with materials	**.60**	.14	.17	−.05
K. Pretending	**.59**	.08	.27	−.03
H. Understanding and expressing feelings	**.58**	−.06	.20	.16
D. Taking care of personal needs	**.51**	.11	.29	−.07
Q. Listening to and understanding speech	**.50**	.06	.07	.31
I. Making and building models	**.43**	.16	.23	.10
2. Cognitive Skills				
V. Using letter names and sounds	−.06	**.98**	−.14	.09
X. Writing	.09	**.89**	−.02	−.07
BB. Counting	−.02	**.66**	.08	.16
Y. Sorting objects	.04	**.52**	.11	.26
W. Reading	.04	**.50**	.05	.27
Z. Identifying patterns	−.01	**.48**	.10	.33
T. Showing awareness of sounds in words	.17	**.47**	.02	.27
J. Drawing and painting pictures	.30	**.41**	.28	−.08
3. Coordinated Movement				
N. Feeling and expressing steady beat	−.03	.01	**.85**	.05
O. Moving to music	.00	−.02	**.81**	.08
P. Singing	.06	−.05	**.66**	.19
L. Moving in various ways	.13	−.04	**.64**	.13
M. Moving with objects	.08	−.02	**.48**	.26
4. Scientific Process Skills				
EE. Identifying materials and properties	−.12	.10	.13	**.80**
FF. Identifying natural and living things	−.08	.01	.15	**.79**
CC. Identifying position and direction	.12	.08	.01	**.68**
DD. Identifying sequence, change, and causality	.15	.10	.02	**.68**
AA. Comparing properties	.07	.20	.11	**.57**
Items not included in final solution				
U. Demonstrating knowledge about books	.30	.28	.10	.23
R. Using vocabulary	**.41**	.14	−.03	**.40**

Note. Salient pattern loadings ($\geq$.40) are in bold. Item R loaded saliently on the Social Engagement and Scientific Process Skills factors and was removed from both. Item titles are from HighScope (2010).

theoretical sense. Based on the loadings, the factors were named Social Engagement (e.g., ‘‘Relating to other children’’), Cognitive Skills (e.g., ‘‘Counting’’), Coordinated Movement (e.g., ‘‘Moving to music’’), and Scientific Process Skills (e.g., ‘‘Identifying natural and living things’’).

CFAs

The confirmatory sample ($n = 2{,}035$) was used to test the fit of the four-factor, 30-item EFA solution using CFA. The CFI (.980) indicated that this model provided a reasonable fit, but the RMSEA (.062) was slightly higher than the established criteria, and the WRMR (1.817) did not meet the criteria for good fit ($\chi^2(399) = 3{,}521.57$, $p < .00001$). Akaike's information criterion (AIC) for the four-factor solution was less than the AIC for the six-factor solution (estimated using the confirmatory sample), indicating that the four-factor model fit the data better (2,239.43 vs. 2,956.95, respectively). The modification indices and SEPC estimates indicated that some items would load saliently on another factor. However, fewer changes were suggested for the four-factor model than for the six-factor model, again supporting the simpler model (8 vs. 19 suggestions, respectively, of salient cross-factor item loadings). Still, high interfactor correlations ($M\ r = .89$, range = .83–.94; see Table 4) called into question the extent to which the four factors represented distinct constructs.

An orthogonalized second-order factor analysis was also performed (Brown, 2006; Watkins et al., 2006). The second-order factor loadings, residualized first-order loadings, variance explained by the second- and first-order factors, and communalities are shown in Table 5. Every COR-2 item loaded saliently on the second-order factor, whereas none of the residualized first-order loadings were salient. Examining the variance explained by the second- and first-order loadings revealed that the second-order factor largely accounted for the variance. Overall, the second-order factor accounted for 64% of the total variance and 90% of the common variance, whereas the first-order factors collectively accounted only for 7% and 10%, respectively. The results suggested that a second-order factor may have accounted for the high interfactor correlations. However, it should be noted that the second-order model did not meet all of the established fit criteria, with a CFI of .98, an RMSEA of .065, and a WRMR of 1.94 ($\chi^2(401) = 3{,}819.99$, $p < .00001$). Still, collectively, these and the previous results raised questions about the utility of the four-factor model.

Developmental Sequence

To further investigate the four-factor model, we tested the individual items for threshold reversals and for reasonable threshold spacing using the full sample ($N = 4{,}071$). The items were examined by each factor because unidimensionality is an assumption of standard IRT models (Embretson & Reise, 2000) and because multidimensionality may contribute to threshold reversals (Andrich et al., 1997). The assumption of unidimensionality and the concern that multidimensionality may be causing reversals were addressed by examining the thresholds by factor.

TABLE 4
Interfactor Correlations for the Four-Factor Model

Factor	*1*	*2*	*3*	*4*
1. Social Engagement	—	.91	.90	.92
2. Cognitive Skills		—	.83	.94
3. Coordinated Movement			—	.87
4. Scientific Process Skills				—

TABLE 5
Orthogonalized Higher Order Model Loadings and Variance Explained

	General status		*Social engagement*		*Cognitive skills*		*Coordinated movement*		*Scientific process*		
Item	*Est*	*%v*	*Est*	*%v*	*Est*	*%v*	*Est*	*%v*	*Est*	*%v*	h^2
E. Relate to adults	.81	.66	.23	.05							.71
F. Relate to children	.77	.60	.22	.05							.64
A. Choices and plans	.80	.64	.23	.05							.69
G. Resolve conflict	.73	.54	.21	.04							.58
C. Initiating play	.82	.67	.23	.05							.72
S. Complex speech	.82	.68	.23	.05							.73
B. Solve problems	.78	.61	.22	.05							.66
K. Pretending	.81	.66	.23	.05							.71
H. Feelings	.80	.64	.23	.05							.69
D. Personal needs	.76	.57	.22	.05							.62
Q. Understand speech	.85	.72	.24	.06							.78
I. Models	.82	.67	.23	.05							.72
V. Letters	.75	.56			.25	.06					.62
X. Writing	.78	.61			.26	.07					.67
BB. Counting	.80	.64			.26	.07					.71
Y. Sorting objects	.85	.71			.28	.08					.79
W. Reading	.76	.58			.25	.06					.64
Z. Identify patterns	.81	.66			.27	.07					.73
T. Sounds in words	.81	.66			.27	.07					.73
J. Making pictures	.80	.63			.26	.07					.70
N. Steady beat	.79	.63					.36	.13			.76
O. Move to music	.78	.61					.36	.13			.74
P. Singing	.76	.58					.35	.12			.71
L. Move various ways	.76	.58					.35	.12			.70
M. Move with objects	.72	.52					.33	.11			.64
EE. Materials and properties	.83	.69							.23	.05	.74
FF. Living things	.81	.66							.22	.05	.70
CC. Position and direction	.83	.69							.23	.05	.74
DD. Sequence and causality	.87	.76							.24	.06	.82
AA. Properties	.86	.74							.23	.05	.79
% Total variance		63.8		2.1		1.8		2.0		0.9	70.6
% Common variance		90.4		2.9		2.6		2.9		1.2	100

Note. Loadings were transformed with the Schmid–Leiman procedure. Item titles are from HighScope (2010). Est = factor loading; %v = percent variance explained; h^2 = communality.

The four factors were each calibrated via the PCM and GPCM. A chi-square difference test of model fit indicated that for each of the four factors the GPCM fit better. Furthermore, for each of the factors, the GPCM yielded higher internal consistency and total test information (calculated as the inverse of test error: $1/SE^2$). Thus, the GPCM was retained to evaluate the item skill points. Table 6 provides the GPCM parameter estimates for each item by factor. The first column, α_i, provides estimates of Item *i*'s discrimination parameter, which indicates the degree to which skill point selection varies by ability level (Embretson & Reise, 2000). For Social Engagement, the item discriminations ranged from 0.86 to 1.29 (M $\alpha = 1.07$). For Cognitive Skills,

TABLE 6
Parameter Estimates From the GPCM ($N = 4{,}071$)

	GPCM parameter estimates						Skill point selection percentages				
Item	α_i	*SE*	δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	*1*	*2*	*3*	*4*	*5*
Social Engagement											
E. Relating to adults[a]	1.14	0.03	−0.64	−1.14	0.12	0.84	16	8	30	26	21
F. Relating to other children	1.07	0.03	−1.06	−0.98	0.01	0.65	12	12	27	25	25
A. Making choices and plans	1.29	0.03	−1.47	−0.74	0.51	1.45	9	19	39	23	10
G. Resolving interpersonal conflict	0.96	0.02	−1.38	−0.05	0.75	1.50	13	33	28	17	9
C. Initiating play	1.17	0.03	−1.54	−1.09	0.48	0.51	7	14	39	16	24
S. Using complex patterns of speech	1.15	0.03	−1.32	−0.51	0.44	0.59	11	22	28	16	23
B. Solving problems with materials	1.02	0.02	−1.58	−0.48	0.32	1.39	9	25	29	25	12
K. Pretending	1.18	0.03	−0.90	−0.70	−0.10	1.35	15	14	23	35	13
H. Understanding and expressing feelings[a]	0.86	0.02	−0.20	−0.63	0.69	0.33	27	13	24	12	23
D. Taking care of personal needs	0.89	0.02	−2.92	−1.15	0.23	0.50	2	16	34	22	26
Q. Listening to and understanding speech	1.11	0.03	−0.58	−0.26	−0.07	0.77	23	16	15	23	22
I. Making and building models	1.02	0.03	−0.85	−0.54	0.38	1.32	18	18	28	23	13
Cognitive Skills											
V. Using letter names and sounds[a]	1.10	0.03	−0.10	0.61	0.40	1.35	40	23	11	16	10
X. Writing[a]	1.23	0.03	−0.16	−0.56	0.89	2.12	30	11	35	19	4
BB. Counting	1.13	0.03	−1.09	−0.26	0.70	1.00	16	25	29	16	15
Y. Sorting objects	1.27	0.03	−0.85	−0.26	0.62	1.39	20	22	29	19	10
W. Reading	1.04	0.03	−1.65	−0.07	0.33	2.35	10	32	25	30	4
Z. Identifying patterns[a]	0.93	0.02	−0.24	−0.09	1.12	0.81	32	20	25	10	14
T. Showing awareness of sounds in words	1.23	0.03	−0.86	0.72	0.73	1.78	23	43	14	14	6
J. Drawing and painting pictures[a]	0.83	0.02	−0.60	−1.27	0.74	1.06	16	10	39	20	15
Coordinated Movement											
N. Feeling and expressing steady beat	1.49	0.04	−1.55	−0.41	0.50	0.66	9	26	28	15	22
O. Moving to music	1.33	0.04	−1.50	−0.19	0.35	0.77	10	29	22	18	21
P. Singing[a]	1.09	0.03	−1.26	−1.58	0.52	0.39	8	7	42	16	27
L. Moving in various ways[a]	0.81	0.02	−2.33	−0.83	0.11	−0.18	4	18	23	16	39
M. Moving with objects	0.67	0.02	−1.13	−1.08	0.72	1.32	12	16	37	22	13
Scientific Process Skills											
EE. Identifying materials and properties	1.35	0.04	−0.27	−0.17	0.87	1.42	32	16	28	15	9
FF. Identifying natural and living things[a]	1.06	0.03	−0.48	−0.10	0.93	0.73	27	21	24	11	16
CC. Identifying position and direction	1.34	0.04	−0.76	−0.02	0.82	2.28	23	26	28	20	3
DD. Identifying sequence, change, and causality[a]	1.58	0.05	−0.50	0.17	1.07	0.89	30	26	23	7	14
AA. Comparing properties	1.34	0.04	−0.52	−0.04	0.41	1.96	27	20	21	27	5

Note. *SE* is the standard error for the slope parameter estimate; δij is the threshold parameter, with each item having four thresholds (j), or one fewer than the number of skill points ($j = m - 1$). Item titles are from HighScope (2010). GPCM = generalized partial credit model.

[a]Item thresholds are disordered.

discriminations ranged from 0.83 to 1.27 (M $\alpha = 1.10$). Item discrimination parameters for Coordinated Movement ranged from 0.67 to 1.49 (M $\alpha = 1.08$). Finally, discrimination ranged from 1.06 to 1.58 (M $\alpha = 1.33$) for Scientific Process Skills.

The δij columns provide the item threshold parameter estimates, which indicate the level of the latent trait that separates two adjacent skill points (Embretson & Reise, 2000). The δij

therefore represent the intersection point of the category characteristic curves (CCCs) for adjacent skill points (Adams et al., 2012). The CCCs show the probabilities of a child being rated at each of the skill points as a function of his or her ability with respect to the underlying latent trait (Embretson & Reise, 2000). Figure 2 shows the CCCs generated by PARSCALE 4.1 (Muraki & Bock, 2002) for Item Y, which has ordered and well-spaced thresholds.

Of the 30 items in the four-factor model, 10 displayed threshold reversals (these items are indicated in Table 6). For example, Figure 3 shows the CCC for Item V, ''Using letter names and sounds,'' on the Cognitive Skills factor. Examining this CCC, it is clear that Thresholds 2 and 3 are reversed. Threshold 2, between Skill Point 2 (naming letters) and Skill Point 3 (making a letter sound), is located at a higher point along the cognitive ability continuum than Threshold 3, which lies between Skill Points 3 and 4 (which is about naming more letters than Skill Point 2). A possible explanation for this reversal was found in the research on early reading skills (e.g., McBride-Chang, 1999; Treiman, Tincoff, Rodriguez, Mouzaki, & Francis, 1998), which suggests that making a letter sound (Skill Point 3) is potentially a more advanced skill than naming letters (Skill Points 2 and 4). Thus, moving from Skill Point 2 to 3 (Threshold 2), naming letters to making a letter sound, may be more difficult than going from Skill Point 3 to 4 (Threshold 3), making a letter sound to naming letters. Because reversed thresholds may result from infrequently selected categories (Adams et al., 2012), the category response percentages are also presented in Table 6. Half of the 10 items with disordered thresholds had one or more skill point that was selected for 10% or less of the sample, but none of the points were never used.

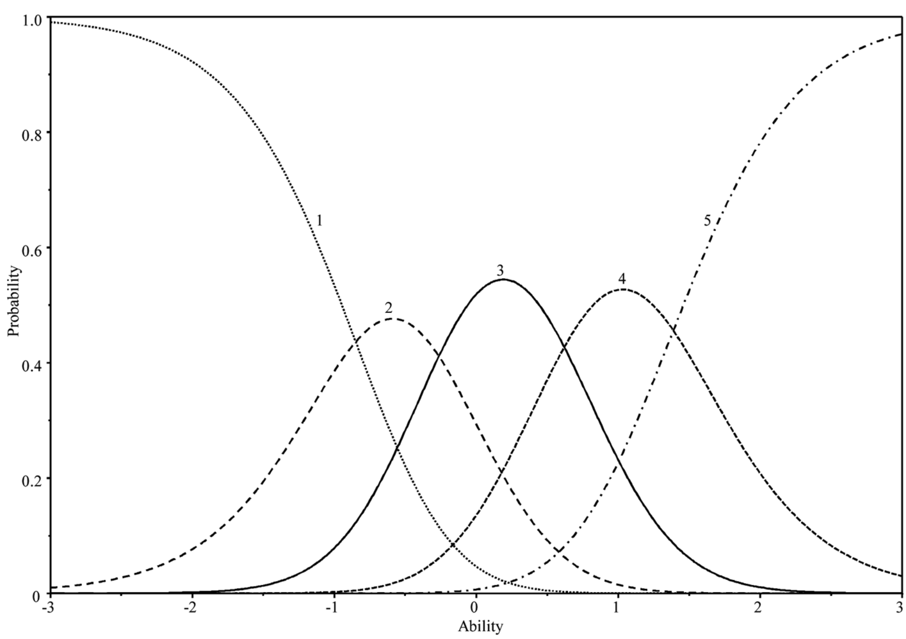

FIGURE 2 Category characteristic curve for Item Y with ordered and well-spaced thresholds. Graphs were generated by PARSCALE 4.1 (Muraki & Bock, 2002).

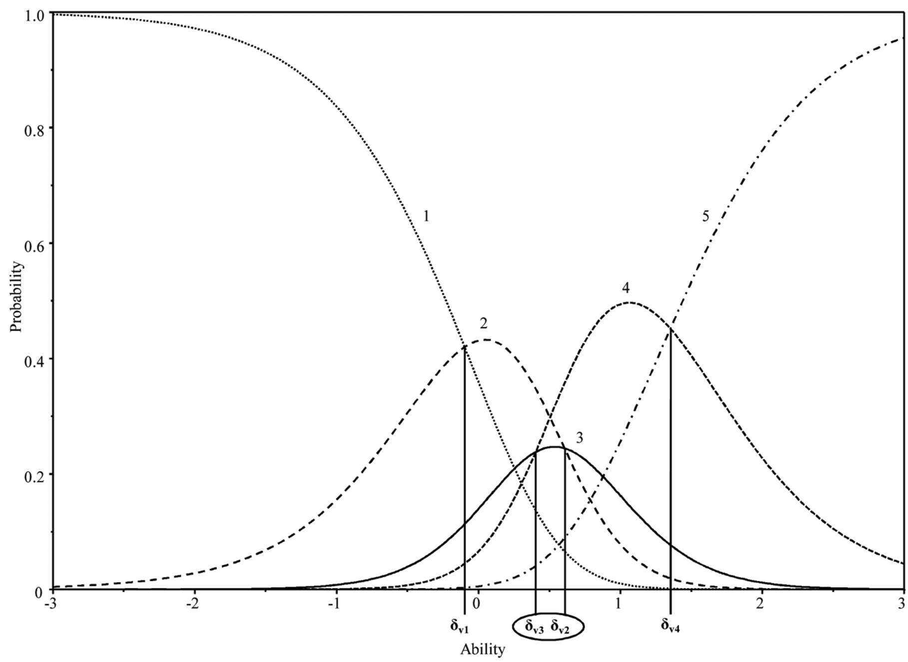

FIGURE 3 Category characteristic curve for Item V with a threshold reversal. δ_{ij} are the threshold parameters. Each item (i) has four thresholds (j). Graphs were generated by PARSCALE 4.1 (Muraki & Bock, 2002). Threshold markers were added.

Finally, an exploratory inspection of the threshold spacing was performed using the results in Table 6. Four of the COR-2 items (C, F, T, and M) had pairs of adjacent thresholds that appeared to be close to one another on the latent trait continuum (i.e., a difference of less than .10). For Item M, ''Moving with objects,'' Threshold 1 (between Skill Points 1 and 2) corresponded to an ability level of -1.13, whereas Threshold 2 (Skill Points 2 and 3) was located at an ability of -1.08 (see Figure 4). Skill Point 2 was the most likely response for only a small portion of the Coordinated Movement continuum. In contrast, the thresholds of Item Y appear to be well spaced, suggesting that each skill point identified a distinct point on the Cognitive Ability continuum (see Figure 2).

As an exploratory and remedial measure, the 10 items with disordered thresholds were removed from their corresponding factors, resulting in two of the four factors retaining only three items. Gorsuch (2003) noted that to robustly define a factor, four or more indicators are needed. Thus, an exploratory investigation of the dimensionality of the 20 items from the four-factor model with ordered thresholds was undertaken. These items were submitted to EFA and CFA to determine the optimal factor structure. Using the exploratory sample, MAP of the smoothed polychoric correlation matrix suggested two potentially viable factors. A three-factor model was also tested, but it did not retain four salient loadings on the third factor. In the two-factor model, one of the items loaded on two of the factors and was removed from both. The two factors were interpreted based on items with the largest loadings, which indicated that they represented

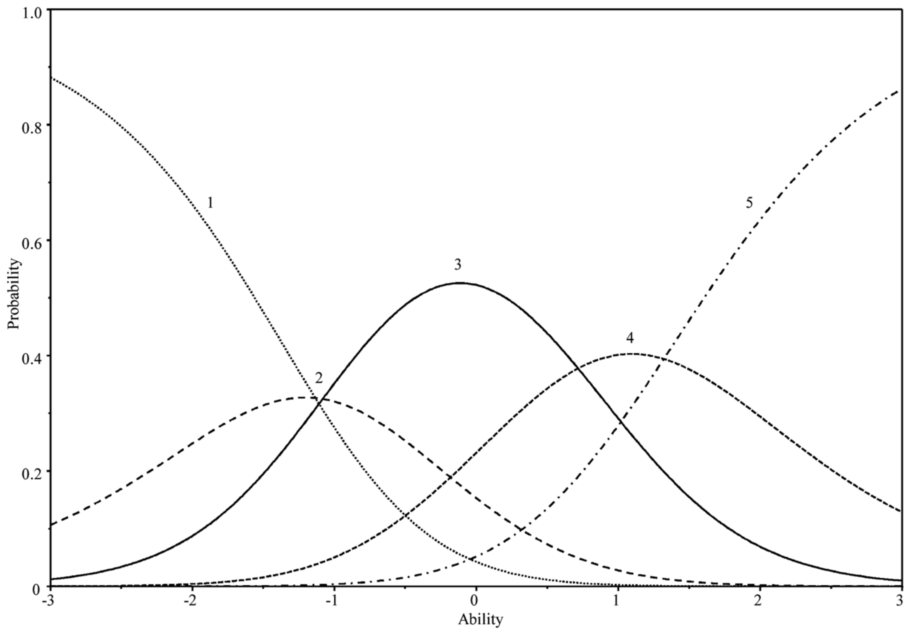

FIGURE 4 Category characteristic curve for Item M with poorly spaced thresholds. Graphs were generated by PARSCALE 4.1 (Muraki & Bock, 2002).

cognitive and noncognitive skills. Using the reserve sample, the CFA of the 19-item two-factor model indicated that this model did not fit well in terms of the criteria set for the global fit indices (CFI = .981, RMSEA = .077, WRMR = 1.909, $\chi^2(151) = 1{,}956.04$, $p < .00001$); in addition, the factors were highly correlated ($r = .93$). Finally, a one-factor model was fit using the 20 items, but its global fit indices were worse than those of the two-factor model.

DISCUSSION

This study provided the first independent investigation of the second most widely used multidimensional assessment in Head Start—the COR-2. The internal structure of the COR-2 was examined by (a) testing the fit of the six developer-defined categories to the data, (b) empirically deriving the optimal factor structure, and (c) testing the hypothesis that the five skill points of each item represented appropriately sequenced and reasonably spaced response options. The modification indices and very high interfactor correlations indicated problems with the model based on the developer-defined categories. EFA, CFA, and high-order factor analysis were used to empirically derive the optimal internal structure. A four-factor model fit the data better than the six-factor model. However, high interfactor correlations were again found, calling into question the extent to which the factors were distinct. An orthogonalized second-order factor analysis suggested that the four first-order factors explained only a small proportion of the variance compared to the second-order factor. Finally, an examination of the skill points of the items indicated that about a third had threshold reversals and four had poorly spaced thresholds.

The only other unpublished study of the COR-2 did not test the fit of the six categories (HighScope, 2010). Thus, currently there is no empirical research available on the COR-2 to support the use of scores based on the six categories. The present study also found very high interfactor correlations for all of the structures assessed. Lower interfactor correlations have been found for other validated early childhood assessments, such as the Learning Express ($M\ r = .66$) and Learning-to-Learn Scales ($M\ r = .61$; McDermott et al., 2009, 2011). However, the Learning Express and Learning-to-Learn Scales may be the exceptions, as highly correlated factors have been found for other widely used early childhood assessments. For example, in an investigation of the Wechsler Intelligence Scale for Children–Fourth Edition, Watkins et al. (2006) found initial support for a four-factor solution with high interfactor correlations. However, an orthogonalized higher order model revealed that a general factor explained the majority of common (76%) and total (47%) variance (the first-order factors collectively explained 15% of common and 24% of total variance; Watkins et al., 2006). Watkins et al. (2006) concluded that given the weak explanatory power of the first-order factors, it would be a mistake to favor their interpretation over the general factor. A similar conclusion could be tentatively drawn from the findings reported here.

The threshold reversal and spacing issues uncovered for the COR-2 are in line with the conclusion of Fantuzzo et al. (2002) for the COR-1 that many of the items have skill points that do not indicate a developmental progression. However, the IRT methods used in the present study to investigate item functioning are not common in evaluations of early childhood assessments (Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2012). Still, this small research base demonstrates that other measures have similar issues with their items functioning. Andrich and Styles (2004) found that many of the items on the Early Development Instrument had disordered thresholds. Gordon et al. (2012) found that all of the items on the Early Childhood Environment Rating Scale–Revised had threshold reversals and about two thirds also had poorly spaced thresholds. Both studies recommended further development using information gleaned from modern psychometric methodologies not available when many of these measures were first developed (Gordon et al., 2012). This recommendation is also applicable for the future development of the COR-2, as discussed next.

Implications for Future Research and Policy

The purpose of this research was to examine the psychometric quality of the COR-2 for use as a multidimensional assessment in Head Start. However, the study is limited by the nature of the Head Start sample and the characteristics of the teachers. This study provided an exploratory investigation of the use of the COR-2 in Head Start in one large, urban school district that was primarily serving an African American population. The data analyzed were collected by Head Start teachers who were required by the district to have at least a bachelor's degree as well as a certification in early childhood education. Teachers' knowledge and skills all influence their ability to effectively use an assessment and help to determine their training needs, such that teachers with fewer qualifications need more training (Mathematica Policy Research, 2007). The results of this study may not be generalizable to teachers with fewer qualifications and/or or different training, to other early childhood programs, or to children from other ethnicities and locales (e.g., rural areas). Additional studies are needed to determine whether the issues uncovered here apply broadly.

This study provides an indication of where the COR-2 is in its scientific development at this time and points to several paths for future research and development. Specifically, the findings presented here highlight the need to further develop the COR-2 in terms of the constructs it aims to measure and the items it uses to do so. Per Downing and Haladyna (2006), an iterative cycle of systematic development and validation should be used for all areas of future work. This iterative cycle involves using information collected during development to inform validation and evidence collected during validation to inform further development (Downing & Haladyna, 2006).

Like many early childhood assessments, the COR-2 aims to measure six constructs essential for school readiness. Measures that only provide information on a subset of these constructs or on general developmental status have less practical utility for practitioners and policymakers. This study's findings do not support the use of scores based on the six categories of the COR-2 and indicate that further development is needed. Future work should use an iterative cycle of construct development and validation, such as the evidence-centered design approach suggested by Mislevy and Riconscente (2006). The construct development phase of evidence-centered design includes working with teachers, experts, and researchers to delineate the facets of the construct (Mislevy & Riconscente, 2006). This process provides a rigorous approach to construct development that yields validity evidence based on content and precise and distinctive construct definitions to be used for item development (Mislevy & Riconscente, 2006). It also helps to ensure that a measure is created that can be used to make valid and reliable inferences about each of the constructs targeted.

For each of the COR-2 items, the skill points purport to capture the developmental sequence of the construct facet represented by the item title (e.g., for Item M, ''Moving with objects'' is the title). However, this study revealed problems with the sequencing and spacing of some of the items' skill points. As the functioning of the COR-2's items was previously unexamined, future research should replicate these results with a large representative sample of children and teachers. Based on this information and the results of this study, problematic items should be flagged for redevelopment using qualitative and quantitative procedures (LeBoeuf, Fantuzzo, & Lopez, 2010). The qualitative investigations would involve having external subject matter experts identify potential sources of malfunctioning (e.g., invalid sequence; Gordon et al., 2012). In addition, teachers could be asked to ''think aloud'' as they observe, take notes, and use their observations to respond to the items (Cook & Beckman, 2006). Findings from this research would identify problems with the response process, such as difficulties in interpreting the skill points (Gordon et al., 2012). The items would then be revised based on the information gathered and knowledge from current developmental theory and research (LeBoeuf et al., 2010). Quantitative investigations using IRT modeling can then empirically confirm that the revised items are functioning as intended.

There are many challenges currently facing Head Start, however, arguably none is more pressing than the need for high-quality assessments (Advisory Committee on Head Start Research and Evaluation, 2012; NRC, 2008). Assessments leading to invalid and unreliable inferences result in decisions with potentially negative consequences for children, teachers, and programs (McDermott et al., 2011). The need for psychometrically sound assessment has received increasing federal attention. The Improving Head Start for School Readiness Act of 2007 contained for the first time the requirement that all programs use scientifically based measures. More recently, the U.S. Department of Education's Race to the Top—Early Learning Challenge called for states

to implement comprehensive early childhood assessment systems that consist of scientifically based measures applicable to the diverse populations served. As the demand for the use of evidence-based assessments increases, so does the need to provide comprehensive information on the quality of measures (NRC, 2008).

There is a significant need to examine what is mandated and the evidence-based capacities of widely used assessments. The NRC (2008) provides an overview of the standards for scientifically based measures as well as a table of widely used assessments. However, this report does not apply the quality standards to the table of assessments, calling for the field to provide the evidence to bridge this gap (LeBoeuf et al., 2010). The present research responded to the NRC's call by investigating the psychometric quality of the second most widely used multidimensional assessment in Head Start—the COR-2. This study is part of a growing body of research (e.g., Fantuzzo, McDermott, Manz-Holliday, Hampton, & Burdick-Alvarez, 1996; LeBoeuf et al., 2010) that investigates early childhood assessments that are widely used with preschool children from low-income households. Without scientific investigation of the quality of assessments, the efficacy of educational programs for young children at high risk for academic failure is in serious jeopardy.

The results of this study also have larger implications for policy. Multidimensional assessments are essential for Head Start to meet its goal of promoting school readiness across important domains discussed in the Child Development and Early Learning Framework, including ''physical well-being, social and emotional development, approaches toward learning, language and literacy skills, and cognitive and general knowledge skills'' (Office of Head Start, 2010, p. 6). To align with Head Start's guiding framework for child outcomes, assessments must be capable of providing scientific evidence on each domain. In addition to providing data on the state of the school readiness domains, Head Start programs are also held accountable for monitoring progress in these areas. For programs to do this, assessments with items that reflect a valid developmental sequence and therefore are able to capture growth are needed. Beyond serving an accountability function, such assessments also improve practice by monitoring children's progress and guiding instructional decisions to create appropriate opportunities for further development. Thus, multidimensional assessments capable of measuring growth provide the actionable intelligence for policymakers and teachers to fulfill Head Start's school readiness goal.

In addressing the pressing need and mandates for scientifically based measures, it is essential that the evidence on the psychometric integrity of early childhood assessments is shared with leaders in early childhood education. These individuals need access to independent evaluations of the assessments they are considering using in their programs. To ensure that psychometric evidence is available and disseminated beyond academic journals and circles, a consumer guide for practitioners and policymakers containing such information is needed.

To move beyond simply mandating the use of scientifically based measures, a national research agenda is needed to ensure that the school readiness of the most vulnerable children is scientifically measured. This agenda should include assessment evaluation, development, and information dissemination. To pursue these important but difficult tasks, the government must make appropriations to financially support these efforts. There is a clear need for scientifically based measures and calls for their use have been made. These calls must be answered with systematic and rigorous science to provide evidence that supports the quality education all young children deserve.

FUNDING

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education (ED), through Grant #R305B090015 to the University of Pennsylvania. The project described was also supported by the Child Care Research Scholars Grant Program, Grant #90YE0138, from the Office of Planning, Research and Evaluation (OPRE), Administration for Children and Families (ACF), U.S. Department of Health and Human Services (DHHS). The contents of the research reported here are solely the responsibility of the authors and do not represent the official views of IES, the U.S. ED, OPRE, ACF, or the U.S. DHHS.

REFERENCES

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*, 547–573. doi:10.1177/0013164411432166

Advisory Committee on Head Start Research and Evaluation. (2012). *Advisory Committee on Head Start Research and Evaluation: Final report*. Retrieved from http://www.acf.hhs.gov/sites/default/files/opre/eval_final.pdf

Alkens, N., Tarullo, L., Hulsey, L., Ross, J., West, J., & Xue, Y. (2010). *ACF-OPRE report: A year in Head Start: Children, families and programs*. Washington, DC: Administration for Children and Families, Office of Planning, Research and Evaluation.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 123–152). New York, NY: Taylor & Francis.

Andrich, D., de Jong, J. H., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59–70). Münster, Germany: Waxmann.

Andrich, D., & Styles, I. (2004). *Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI)*. Retrieved from http://www.rch.org.au/emplibrary/australianedi/Final_Rasch_report.pdf

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105*, 467–477. doi:10.1037/0033-2909.105.3.467

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361–379). New York, NY: Guilford Press.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43. doi:10.1037/a0026975

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754–761. doi:10.1037/0022-006X.56.5.754

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, *119*, 166.e7–166.e16. doi:10.1016/j.amjmed.2005.10.036

Denton Flanagan, K., McPhee, C., & Mulligan, G. (2009). *The children born in 2001 at kindergarten entry: First findings from the kindergarten data collections of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)* (Publication No. NCES 2010005). Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010005

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837. doi:10.1046/j.1365-2923.2003.01594.x

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*, 61–82. doi:10.1207/s15324818ame1001_4

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Epstein, A. S. (1993). *Training for quality: Improving early childhood programs through systematic in-service training* (Monographs of the HighScope Educational Research Foundation No. 9). Ypsilanti, MI: HighScope Press.

Epstein, A. S. (2006). *HighScope and Head Start: A good fit: Forty years of commitment and compatibility*. Retrieved from http://www.highscope.org/file/NewsandInformation/ReSourceReprints/Spring06pdfs/AGoodFit.pdf

Epstein, A. S., & Schweinhart, L. J. (2009). The HighScope preschool curriculum and dimensions of preschool curriculum decision-making. *Early Childhood Services*, *3*, 193–208.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. doi:10.1037/1082-989X.4.3.272

Fantuzzo, J. W., Gadsden, V., & McDermott, P. (2010). *2010 EPIC research brief: Evidence-based Program for the Integration of Curricula (EPIC)*. Retrieved from http://www.gse.upenn.edu/pdf/pennchild/EPIC.pdf

Fantuzzo, J. W., Hightower, D., Grim, S., & Montes, G. (2002). Generalization of the Child Observation Record: A validity study for diverse samples of urban, low-income preschool children. *Early Childhood Research Quarterly*, *17*, 106–125. doi:10.1016/S0885-2006(02)00131-X

Fantuzzo, J. W., McDermott, P. A., Manz-Holliday, P., Hampton, V. R., & Burdick-Alvarez, N. (1996). The Pictorial Scale of Perceived Competence and Social Acceptance: Does it work with low-income urban children? *Child Development*, *67*, 1071–1084. doi:10.2307/1131880

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491. doi:10.1037/1082-989X.9.4.466

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2012). An assessment of the validity of the ECERS–R with implications for measures of child care quality and relations to child development. *Developmental Psychology*. Advance online publication. doi:10.1037/a0027899

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 143–164). New York, NY: Wiley.

Halle, T., Zaslow, M., Wessel, J., Moodie, S., & Darling-Churchill, K. (2011). *Understanding and choosing assessments and developmental screeners for young children: Profiles of selected measures* (OPRE Report No. 2011–23). Retrieved from http://www.acf.hhs.gov/programs/opre/hs/dev_screeners/reports/screeners_final.pdf

HighScope Educational Research Foundation. (1992). *HighScope Child Observation Record (COR) for ages 2½–6*. Ypsilanti, MI: HighScope Press.

HighScope Educational Research Foundation. (2003). *What's different about the new Preschool Child Observation Record (COR)?* Retrieved from http://www.highscope.org/file/Assessment/Whatsdiff2.pdf

HighScope Educational Research Foundation. (2010). *Preschool Child Observation Record, 2nd edition: User guide*. Ypsilanti, MI: HighScope Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

Improving Head Start for School Readiness Act of 2007, P.L. 110–134, 42 U.S.C. § 9801 et seq (2007).

Kane, M. T. (2006a). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, NJ: Erlbaum.

Kane, M. T. (2006b). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Knol, D. L., & Berger, M. P. F. (1988). *Empirical comparison between factor analysis and item response models* (Research Report No. 88–11). Enschede, The Netherlands: University of Twente, Department of Education.

LeBoeuf, W. A., Fantuzzo, J. W., & Lopez, M. L. (2010). Measurement and population miss-fits: A case study on the importance of using appropriate measures to evaluate early childhood interventions. *Applied Developmental Science*, *14*, 45–53. doi:10.1080/10888690903510349

Mathematica Policy Research. (2007). *Measuring children's progress from preschool through third grade*. Retrieved from http://www.pewtrusts.org/uploadedFiles/wwwpewtrustsorg/Reports/Pre-k_education/atkins-burnett%20final%20apper%207-3-07.pdf

McBride-Chang, C. (1999). The ABC's of the ABC's: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quarterly*, *45*, 285–308.

McDermott, P. A., Fantuzzo, J. W., Warley, H. P., Waterman, C., Angelo, L. E., Gadsden, V. L., & Sekino, Y. (2011). Multidimensionality of teachers' graded responses for preschoolers' stylistic learning behavior: The Learning-to-Learn Scales. *Educational and Psychological Measurement*, *71*, 148–169. doi:10.1177/0013164410387351

McDermott, P. A., Fantuzzo, J. W., Waterman, C., Angelo, L. E., Warley, H. P., Gadsden, V. L., & Zhang, X. (2009). Measuring preschool cognitive growth while it's still happening: The Learning Express. *Journal of School Psychology*, *47*, 337–366. doi:10.1016/j.jsp.2009.07.002

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Muraki, E., & Bock, D. (2002). *PARSCLE* (Version 4.1) [Computer software]. Chicago, IL: Scientific Software International.

Muthén, B. O., & Muthén, L. K. (2010). *Mplus* (Version 6.1) [Computer software]. Los Angeles, CA: Author.

National Research Council. (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Neill, P. (2004). A better way to do preschool assessment: Announcing the Revised Preschool COR. *HighScope ReSource*, *23*. Retrieved from http://www.highscope.org/file/NewsandInformation/ReSourceReprints/CORarticle.pdf

Office of Head Start. (2010). *The Head Start child development and early learning framework promoting positive outcomes in early childhood programs serving children 3–5 years old: Revised*. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/Assessment/Child%20Outcomes/HS_Revised_Child_Outcomes_Framework%28rev-Sept2011%29.pdf

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61. doi:10.1007/BF02289209

Schweinhart, L. J., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The HighScope Child Observation Record study. *Educational and Psychological Measurement*, *53*, 445–455. doi:10.1177/0013164493053002014

Sekino, Y., & Fantuzzo, J. W. (2005). Validity of the Child Observation Record: An investigation of the relationship between COR dimensions and social-emotional and cognitive outcomes for Head Start children. *Journal of Psychoeducational Assessment*, *23*, 242–260. doi:10.1177/073428290502300304

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*, 148–154. doi:10.1037/0033-2909.106.1.148

Treiman, R., Tincoff, R., Rodriguez, K., Mouzaki, A., & Francis, D. J. (1998). The foundations of literacy: Learning the sounds of letters. *Child Development*, *69*, 1524–1540. doi:10.1111/j.1467-8624.1998.tb06175.x

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327. doi:10.1007/BF02293557

Waller, N. G. (2001). *MicroFACT: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems* (Version 2.0) [Computer software]. St. Paul, MN: Assessment Systems.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education: Or whose score is it anyway? *Early Childhood Research Quarterly*, *27*(1), 46–54. doi:10.1016/j.ecresq.2011.06.003

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scale for Children–Fourth Edition among referred students. *Educational and Psychological Measurement*, *66*, 975–983. doi:10.1177/0013164406288168

Williams, L. J., Ford, L. R., & Nguyen, N. (2002). Basic and advanced measurement models for confirmatory factor analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 366–389). Malden, MA: Blackwell.

Yates, A. (1987). *Multivariate exploratory data analysis: A prospective on exploratory factor analysis*. Albany: State University of New York Press.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California). Retrieved from http://www.statmodel.com/download/Yudissertation.pdf

Zigler, E., & Styfco, S. (2010). *The hidden history of Head Start*. New York, NY: Oxford University Press.